Naive Bayes: Probabilistic Classification

Nipun Batra

IIT Gandhinagar

July 31, 2025

Outline

From Deterministic to Probabilistic

So far: Decision trees, SVM, etc. give hard classifications

From Deterministic to Probabilistic

So far: Decision trees, SVM, etc. give hard classifications
Now: What if we want probabilities?

- "Email is 85% likely to be spam"

From Deterministic to Probabilistic

So far: Decision trees, SVM, etc. give hard classifications
Now: What if we want probabilities?

- "Email is 85% likely to be spam"
- "Patient has 60% chance of having disease"

From Deterministic to Probabilistic

So far: Decision trees, SVM, etc. give hard classifications
Now: What if we want probabilities?

- "Email is 85% likely to be spam"
- "Patient has 60% chance of having disease"
- Confidence in our predictions

From Deterministic to Probabilistic

So far: Decision trees, SVM, etc. give hard classifications
Now: What if we want probabilities?

- "Email is 85% likely to be spam"
- "Patient has 60% chance of having disease"
- Confidence in our predictions

From Deterministic to Probabilistic

So far: Decision trees, SVM, etc. give hard classifications
Now: What if we want probabilities?

- "Email is 85% likely to be spam"
- "Patient has 60% chance of having disease"
- Confidence in our predictions

Naive Bayes Approach

Use **Bayes' Theorem** + **Independence Assumption** for probabilistic classification

Bayes' Theorem Refresher

Bayes' Theorem

$$P(C|X) = \frac{P(X|C) \cdot P(C)}{P(X)}$$

Bayes' Theorem Refresher

Bayes' Theorem

$$P(C|X) = \frac{P(X|C) \cdot P(C)}{P(X)}$$

Bayes' Theorem Refresher

Bayes' Theorem

$$P(C|X) = \frac{P(X|C) \cdot P(C)}{P(X)}$$

For classification:

- $P(C|X)$: **Posterior** - probability of class given features

Bayes' Theorem Refresher

Bayes' Theorem

$$P(C|X) = \frac{P(X|C) \cdot P(C)}{P(X)}$$

For classification:

- $P(C|X)$: **Posterior** - probability of class given features
- $P(X|C)$: **Likelihood** - probability of features given class

Bayes' Theorem Refresher

Bayes' Theorem

$$P(C|X) = \frac{P(X|C) \cdot P(C)}{P(X)}$$

For classification:

- $P(C|X)$: **Posterior** - probability of class given features
- $P(X|C)$: **Likelihood** - probability of features given class
- $P(C)$: **Prior** - probability of class

Bayes' Theorem Refresher

Bayes' Theorem

$$P(C|X) = \frac{P(X|C) \cdot P(C)}{P(X)}$$

For classification:

- $P(C|X)$: **Posterior** - probability of class given features
- $P(X|C)$: **Likelihood** - probability of features given class
- $P(C)$: **Prior** - probability of class
- $P(X)$: **Evidence** - probability of features

Pop Quiz: Bayes' Theorem

Quick Quiz 1

In email spam classification, what does $P(\text{spam}|\text{contains "free"})$ represent?

a) Probability that word "free" appears in spam emails

Answer: b) It's the posterior probability - what we want to predict!

Pop Quiz: Bayes' Theorem

Quick Quiz 1

In email spam classification, what does $P(\text{spam}|\text{contains "free"})$ represent?

a) Probability that word "free" appears in spam emails
b) Probability that email is spam given it contains "free"

Answer: b) It's the posterior probability - what we want to predict!

Pop Quiz: Bayes' Theorem

Quick Quiz 1

In email spam classification, what does $P(\text{spam}|\text{contains "free"})$ represent?

a) Probability that word "free" appears in spam emails
b) Probability that email is spam given it contains "free"
c) Probability that any email contains "free"

Answer: b) It's the posterior probability - what we want to predict!

Why "Naive"? The Independence Assumption

Problem: For multiple features $X = (x_1, x_2, \ldots, x_n)$:

$$P(X|C) = P(x_1, x_2, \ldots, x_n|C)$$

This joint probability is hard to estimate!

Why "Naive"? The Independence Assumption

Problem: For multiple features $X = (x_1, x_2, \ldots, x_n)$:

$$P(X|C) = P(x_1, x_2, \ldots, x_n|C)$$

This joint probability is hard to estimate!

Naive Assumption

Assume features are conditionally independent given class:

$$P(x_1, x_2, \ldots, x_n|C) = P(x_1|C) \cdot P(x_2|C) \cdots P(x_n|C)$$

Why "Naive"? The Independence Assumption

Problem: For multiple features $X = (x_1, x_2, \ldots, x_n)$:

$$P(X|C) = P(x_1, x_2, \ldots, x_n|C)$$

This joint probability is hard to estimate!

Naive Assumption

Assume features are conditionally independent given class:

$$P(x_1, x_2, \ldots, x_n|C) = P(x_1|C) \cdot P(x_2|C) \cdots P(x_n|C)$$

Why "Naive"? The Independence Assumption

Problem: For multiple features $X = (x_1, x_2, \ldots, x_n)$:

$$P(X|C) = P(x_1, x_2, \ldots, x_n|C)$$

This joint probability is hard to estimate!

Naive Assumption

Assume features are conditionally independent given class:

$$P(x_1, x_2, \ldots, x_n|C) = P(x_1|C) \cdot P(x_2|C) \cdots P(x_n|C)$$

Why "naive"? This assumption is often violated in real data, but works surprisingly well!

When Features are Continuous: Gaussian NB

For continuous features: Assume each feature follows a Gaussian distribution

Gaussian Assumption

$$P(x_i|C) = \frac{1}{\sqrt{2\pi\sigma_{i,C}^2}} \exp\left(-\frac{(x_i - \mu_{i,C})^2}{2\sigma_{i,C}^2}\right)$$

When Features are Continuous: Gaussian NB

For continuous features: Assume each feature follows a Gaussian distribution

Gaussian Assumption

$$P(x_i|C) = \frac{1}{\sqrt{2\pi\sigma_{i,C}^2}} \exp\left(-\frac{(x_i - \mu_{i,C})^2}{2\sigma_{i,C}^2}\right)$$

When Features are Continuous: Gaussian NB

For continuous features: Assume each feature follows a Gaussian distribution

Gaussian Assumption

$$P(x_i|C) = \frac{1}{\sqrt{2\pi\sigma_{i,C}^2}} \exp\left(-\frac{(x_i - \mu_{i,C})^2}{2\sigma_{i,C}^2}\right)$$

Parameters to learn:

- $\mu_{i,C}$: Mean of feature i for class C

When Features are Continuous: Gaussian NB

For continuous features: Assume each feature follows a Gaussian distribution

Gaussian Assumption

$$P(x_i|C) = \frac{1}{\sqrt{2\pi\sigma_{i,C}^2}} \exp\left(-\frac{(x_i - \mu_{i,C})^2}{2\sigma_{i,C}^2}\right)$$

Parameters to learn:

- $\mu_{i,C}$: Mean of feature *i* for class *C*
- $\sigma_{i,C}^2$: Variance of feature *i* for class *C*

Important Considerations

1. Why is Naive Bayes particularly effective for text classification?

Important Considerations

1. Why is Naive Bayes particularly effective for text classification?

Important Considerations

1. Why is Naive Bayes particularly effective for text classification?
2. What happens when a feature value appears in test data but not in training data?

Important Considerations

1. Why is Naive Bayes particularly effective for text classification?
2. What happens when a feature value appears in test data but not in training data?

Important Considerations

1. Why is Naive Bayes particularly effective for text classification?
2. What happens when a feature value appears in test data but not in training data?
3. Compare Naive Bayes with logistic regression - when would you choose each?

Key Takeaways

- **Probabilistic Foundation**: Based on Bayes′ theorem and conditional independence

Key Takeaways

- **Probabilistic Foundation**: Based on Bayes′ theorem and conditional independence

Key Takeaways

- **Probabilistic Foundation**: Based on Bayes′ theorem and conditional independence
- **Naive Assumption**: Features are conditionally independent given the class

Key Takeaways

- **Probabilistic Foundation**: Based on Bayes′ theorem and conditional independence
- **Naive Assumption**: Features are conditionally independent given the class

Key Takeaways

- **Probabilistic Foundation**: Based on Bayes′ theorem and conditional independence
- **Naive Assumption**: Features are conditionally independent given the class
- **Efficient Training**: Simple parameter estimation from training data

Key Takeaways

- **Probabilistic Foundation**: Based on Bayes′ theorem and conditional independence
- **Naive Assumption**: Features are conditionally independent given the class
- **Efficient Training**: Simple parameter estimation from training data

Key Takeaways

- **Probabilistic Foundation**: Based on Bayes′ theorem and conditional independence
- **Naive Assumption**: Features are conditionally independent given the class
- **Efficient Training**: Simple parameter estimation from training data
- **Handles Multiple Classes**: Naturally extends to multi-class problems

Key Takeaways

- **Probabilistic Foundation**: Based on Bayes′ theorem and conditional independence
- **Naive Assumption**: Features are conditionally independent given the class
- **Efficient Training**: Simple parameter estimation from training data
- **Handles Multiple Classes**: Naturally extends to multi-class problems

Key Takeaways

- **Probabilistic Foundation**: Based on Bayes′ theorem and conditional independence
- **Naive Assumption**: Features are conditionally independent given the class
- **Efficient Training**: Simple parameter estimation from training data
- **Handles Multiple Classes**: Naturally extends to multi-class problems
- **Good with Small Data**: Works well with limited training examples

Key Takeaways

- **Probabilistic Foundation**: Based on Bayes′ theorem and conditional independence
- **Naive Assumption**: Features are conditionally independent given the class
- **Efficient Training**: Simple parameter estimation from training data
- **Handles Multiple Classes**: Naturally extends to multi-class problems
- **Good with Small Data**: Works well with limited training examples

Key Takeaways

- **Probabilistic Foundation**: Based on Bayes′ theorem and conditional independence
- **Naive Assumption**: Features are conditionally independent given the class
- **Efficient Training**: Simple parameter estimation from training data
- **Handles Multiple Classes**: Naturally extends to multi-class problems
- **Good with Small Data**: Works well with limited training examples
- **Interpretable**: Probabilistic outputs provide confidence measures